Random sampling locations for comparing a mean with a fixed threshold (parametric)

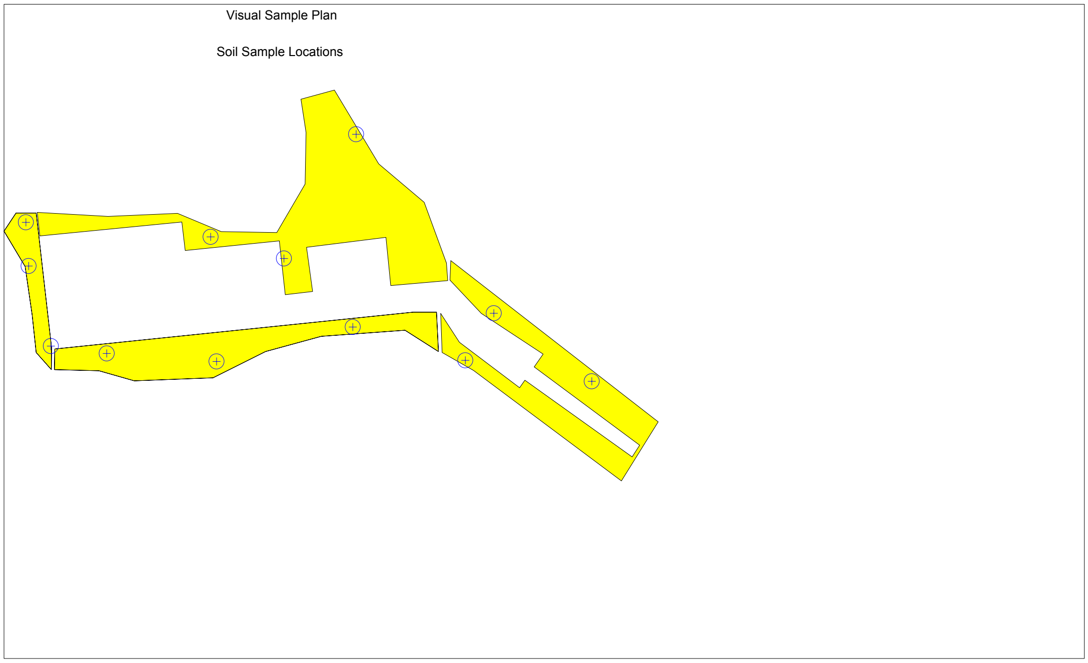
Summary

This report summarizes the sampling design used, associated statistical assumptions, as well as general guidelines for conducting post-sampling data analysis. Sampling plan components presented here include how many sampling locations to choose and where within the sampling area to collect those samples. The type of medium to sample (i.e., soil, groundwater, etc.) and how to analyze the samples (in-situ, fixed laboratory, etc.) are addressed in other sections of the sampling plan.

The following table summarizes the sampling design developed. A figure that shows sampling locations in the field and a table that lists sampling location coordinates are also provided below.

SUMMARY OF SAMPLING DESIGN	
Primary Objective of Design	Compare a site mean to a fixed threshold
Type of Sampling Design	Parametric
Sample Placement (Location) in the Field	Simple random sampling
Working (Null) Hypothesis	The mean value at the site exceeds the threshold
Formula for calculating number of sampling locations	Student's t-test
Calculated total number of samples	3
Number of samples on map [a]	12
Number of selected sample areas [b]	4
Specified sampling area [c]	5252.76 ft^2
Total cost of sampling [d]	$2500.00

[a] This number may differ from the calculated number because of 1) grid edge effects, 2) adding judgment samples, or 3) selecting or unselecting sample areas.
[b] The number of selected sample areas is the number of colored areas on the map of the site. These sample areas contain the locations where samples are collected.
[c] The sampling area is the total surface area of the selected colored sample areas on the map of the site.
[d] Including measurement analyses and fixed overhead costs. See the Cost of Sampling section for an explanation of the costs presented here.



Area 1

X Coord	Y Coord	Label	Value	Type	Historical
611594.0011	306328.4626		0	Random	
611627.1204	306305.4656		0	Random	
611584.3721	306312.5504		0	Random	

Area 2					
X Coord	Y Coord	Label	Value	Type	Historical
611547.4288	306388.9626		0	Random	
611498.2397	306354.2599		0	Random	
611523.0212	306346.9781		0	Random	

Area 3					
X Coord	Y Coord	Label	Value	Type	Historical
611435.7605	306359.1912		0	Random	
611436.6710	306344.4248		0	Random	
611444.2419	306317.3629		0	Random	

Area 4					
X Coord	Y Coord	Label	Value	Type	Historical
611546.2855	306323.8171		0	Random	
611500.2680	306312.1823		0	Random	
611463.1319	306314.8721		0	Random	

Primary Sampling Objective
The primary purpose of sampling at this site is to compare a mean value with a fixed threshold. The working hypothesis (or 'null' hypothesis) is that the mean value at the site is equal to or exceeds the threshold. The alternative hypothesis is that the mean value is less than the threshold. VSP calculates the number of samples required to reject the null hypothesis in favor of the alternative one, given a selected sampling approach and inputs to the associated equation.

Selected Sampling Approach
A parametric random sampling approach was used to determine the number of samples and to specify sampling locations. A parametric formula was chosen because the conceptual model and historical information (e.g., historical data from this site or a very similar site) indicate that parametric assumptions are true.These assumptions will be examined in post-sampling data analysis.

Both parametric and non-parametric equations rely on assumptions about the population. Typically, however, non-parametric equations require fewer assumptions and allow for more uncertainty about the statistical distribution of values at the site. The trade-off is that if the parametric assumptions are valid, the required number of samples is usually less than if a non-parametric equation was used.

Locating the sample points randomly provides data that are separated by many distances, whereas systematic samples are all equidistant apart. Therefore, random sampling provides more information about the spatial structure of the potential contamination than systematic sampling does. As with systematic sampling, random sampling also provides information regarding the mean value, but there is the possibility that areas of the site will not be represented with the same frequency as if uniform grid sampling were performed.

Number of Total Samples: Calculation Equation and Inputs
The equation used to calculate the number of samples is based on a Student's t-test. For this site, the null hypothesis is rejected in favor of the alternative one if the sample mean is sufficiently smaller than the threshold. The number of samples to collect is calculated so that if the inputs to the equation are true, the calculated number of samples will cause the null hypothesis to be rejected.

The formula used to calculate the number of samples is:

$$n = \frac{S^2}{\Delta^2}\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 + 0.5Z_{1-\alpha}^2$$

where

n is the number of samples,
S is the estimated standard deviation of the measured values including analytical error,
Δ is the width of the gray region,
α is the acceptable probability of incorrectly concluding the site mean is less than the threshold,
β is the acceptable probability of incorrectly concluding the site mean exceeds the threshold,
$Z_{1-\alpha}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\alpha}$ is $1-\alpha$,
$Z_{1-\beta}$ is the value of the standard normal distribution such that the proportion of the distribution less than $Z_{1-\beta}$ is $1-\beta$.

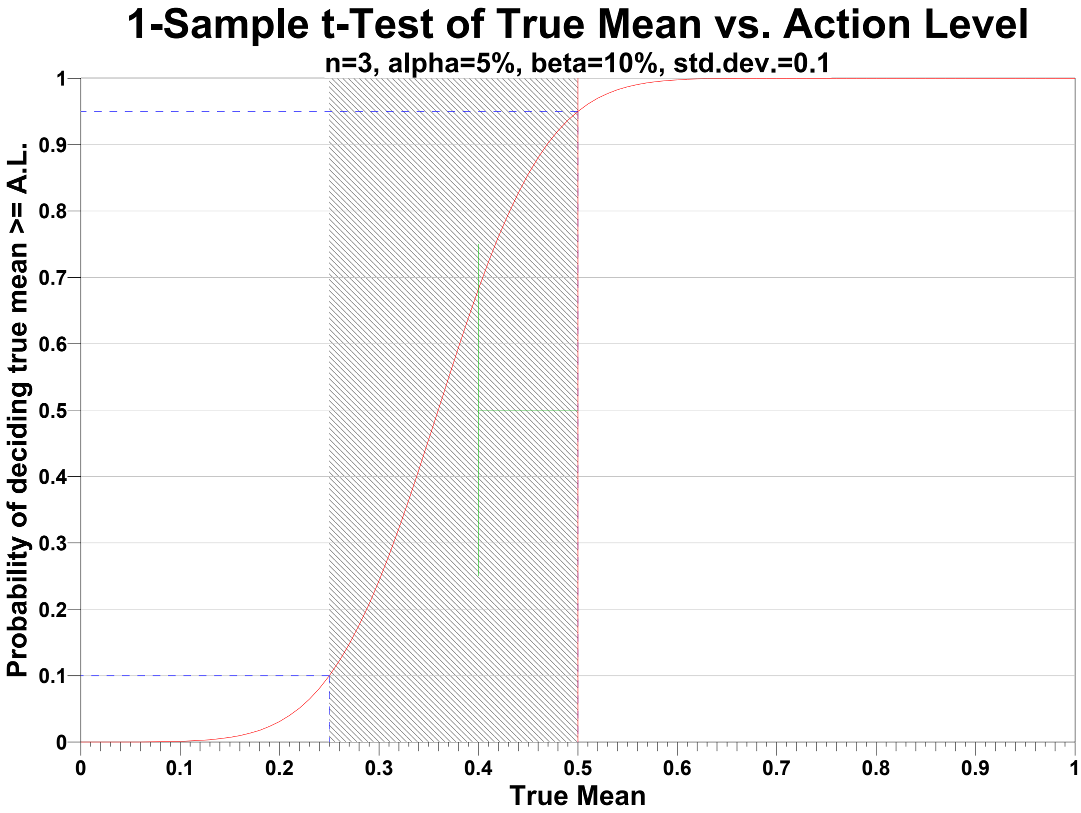
The values of these inputs that result in the calculated number of sampling locations are:

Parameter	**Value**
S	0.1
Δ	0.25
α	5%
β	10%
$Z_{1-\alpha}$	1.64485 [a]
$Z_{1-\beta}$	1.28155 [b]

[a] This value is automatically calculated by VSP based upon the user defined value of α.
[b] This value is automatically calculated by VSP based upon the user defined value of β.

The following figure is a performance goal diagram, described in EPA's QA/G-4 guidance (EPA, 2000). It shows the probability of concluding the sample area is dirty on the vertical axis versus a range of possible true mean values for the site on the horizontal axis. This graph contains all of the inputs to the number of samples equation and pictorially represents the calculation.

The red vertical line is shown at the threshold (action limit) on the horizontal axis. The width of the gray shaded area is equal to Δ; the upper horizontal dashed blue line is positioned at $1-\alpha$ on the vertical axis; the lower horizontal dashed blue line is positioned at β on the vertical axis. The vertical green line is positioned at one standard deviation below the threshold. The shape of the red curve corresponds to the estimates of variability. The calculated number of samples results in the curve that passes through the lower bound of Δ at β and the upper bound of Δ at $1-\alpha$. If any of the inputs change, the number of samples that result in the correct curve changes.



Statistical Assumptions

The assumptions associated with the formulas for computing the number of samples are:

1. the sample mean is normally distributed,
2. the variance estimate, S^2, is reasonable and representative of the population being sampled,
3. the population values are not spatially or temporally correlated, and
4. the sampling locations will be selected randomly.

The first three assumptions will be assessed in a post data collection analysis. The last assumption is valid because the sample locations were selected using a random process.

Sensitivity Analysis

The sensitivity of the calculation of number of samples was explored by varying s, LBGR, β and α and examining the resulting changes in the number of samples. The following table shows the results of this analysis.

Number of Samples							
AL=0.5		α=5		α=10		α=15	
		s=0.2	s=0.1	s=0.2	s=0.1	s=0.2	s=0.1
LBGR=90	β=5	175	45	138	36	116	30
	β=10	139	36	106	28	87	23
	β=15	117	31	87	23	70	18
LBGR=80	β=5	45	13	36	10	30	8
	β=10	36	10	28	8	23	6
	β=15	31	9	23	7	18	5
LBGR=70	β=5	21	7	17	5	14	4
	β=10	17	6	13	4	11	3
	β=15	15	5	11	4	9	3

s = Standard Deviation
LBGR = Lower Bound of Gray Region (% of Action Level)
β = Beta (%), Probability of mistakenly concluding that μ > action level

α = Alpha (%), Probability of mistakenly concluding that μ < action level
AL = Action Level (Threshold)

Cost of Sampling

The total cost of the completed sampling program depends on several cost inputs, some of which are fixed, and others that are based on the number of samples collected and measured. Based on the numbers of samples determined above, the estimated total cost of sampling and analysis at this site is $2500.00, which averages out to a per sample cost of $833.33. The following table summarizes the inputs and resulting cost estimates.

COST INFORMATION			
Cost Details	**Per Analysis**	**Per Sample**	**3 Samples**
Field collection costs		$100.00	$300.00
Analytical costs	$400.00	$400.00	$1200.00
Sum of Field & Analytical costs		**$500.00**	**$1500.00**
Fixed planning and validation costs			$1000.00
Total cost			**$2500.00**

Recommended Data Analysis Activities

Post data collection activities generally follow those outlined in EPA's Guidance for Data Quality Assessment (EPA, 2000). The data analysts will become familiar with the context of the problem and goals for data collection and assessment. The data will be verified and validated before being subjected to statistical or other analyses. Graphical and analytical tools will be used to verify to the extent possible the assumptions of any statistical analyses that are performed as well as to achieve a general understanding of the data. The data will be assessed to determine whether they are adequate in both quality and quantity to support the primary objective of sampling.

Because the primary objective for sampling for this site is to compare the site mean value with a threshold value, the data will be assessed in this context. Assuming the data are adequate, at least one statistical test will be done to perform a comparison between the data and the threshold of interest. Results of the exploratory and quantitative assessments of the data will be reported, along with conclusions that may be supported by them.